LA-UR -91-2704

LA-UR--91-2704

DE91 018036

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE WHAT'S HAPPENING WITH SUPERCOMPUTER NETWORKS

AUTHOR(S) DON E. TOLMIE

SUBMITTED TO NETWORK SYSTEMS USERS GROUP ANNUAL CONFERENCE XXIV

DISCLAIMER

MASTER

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

What's Happening with Supercomputer Networks

Don E. Tolmie,

Los Alamos National Laboratory
Los Alamos, New Mexico
(505) 667-5502

ABSTRACT

Computer networks must become faster as the equipment that is being interconnected increases in power and performance. Ethernet, with a 10 Mbit/s speed, seemed awesome a few years ago, but is beginning to show its age as more machines are tied together, and workstations attain the power of yesterdays mainframes.

Networks using gigabit speeds are just starting to become available and offer a whole new set of problems and potential. The networks proposed for supercomputers today will be the run-of-the-mill networks interconnecting workstations and other ADP equipment in the near future. This paper addresses what the higher speeds are being used for, the "standards" efforts specifying the higher speed channels, the network architectures being proposed, and some of the open problems requiring extensive further work.

WHY DO WE NEED GIGABIT NETWORKS

When networks were mainly used to carry key strokes between dumb terminals and mainframes, 9600 baud was quite adequate; it was considerably faster than people could read. Today it is more common to pass files and pictures between the workstations, mainframes, and storage systems. The emphasis is on improving the users productivity and avoiding network bottlenecks.

Visualization

If a picture is worth a thousand words, then remember that it probably also takes a thousand times the bandwidth to transfer that picture. People are not content with just pictures, presenting the computer output data in movie format (called visualization) is the newest craze and offers even higher user productivity increases. The potential bandwidth of the human eye-brain system has been calculated to be on the order of a few gigabits per second, hence gigabit speeds should satisfy the individual user's needs for a while.

The networking factors of importance for visualization are raw speed and non-interference between data streams - if a visualization data stream is interrupted by another packet, then the user sees a glitch which is very distracting. Visualization sessions also tend to last for many seconds, compared to a single packet transfer which may only take a few microseconds. Error control is also unique in that data in error is discarded rather than being retransmitted.

File Transfers

As the computers become faster, they also increase their appetite for data. A computer that is constipated because of bottlenecks for input or output data is wasting useful compute cycles. A major factor is the bandwidth between the computer and its mass storage system. Mass storage systems used to be limited to single disks attached intimately to individual computer systems; today the trend is for groups of disks to be shared among a group of networked workstations. The networking factors of importance for file transfers are raw speed and fairly large files; latency and interfering data streams are not major concerns.

Remote procedure calls

An interesting concept that is gaining acceptance is the close coupling of many workstations to achieve the compute power of a supercomputer. Single CPU supercomputers are running out of potential performance gains due to the laws of physics limiting the speed of light and electrons. Performance gains in the future will be achieved by interconnecting many smaller computers and spreading the problem across all of them. This has been termed "the attack of the killer micros". The networking factors of importance for

remote procedure calls (RPC's) are raw speed, low cost (it shouldn't cost more than the workstation), and low latency. The information transferred tends to be mainly short data, control, and synchronizing packets.

STANDARDS

The computing industry has become aware that hardware and software standards are necessary for future growth. No single company can provide all of the solutions, and interoperation with other vendors requires agreed upon interfaces. The users are also demanding conformance to standards so that they can purchase from multiple vendors, and minimize their training costs.

Some years ago some people thought that standards stifled creativity. It is our observation that standards allow a company to invest a larger amount in their own areas of special expertise, with a smaller investment required to interface to the other vendors that conform to the standard. Otherwise, the cost of separate interfaces to each individual vendor may well outweigh the cost of the main business.

We have also seen that the standards process usually brings together the best and brightest people of many companies to work collectively on a problem. Design by committee really does work; the output of a standards committee is usually considerably more thorough and of higher quality than if one person or one company had done the complete job. We cannot say enough good things about the companies and individuals that support the voluntary standards efforts.

In the gigabit computer networking arena, the High-Performance Parallel Interface (HIPPI) and Fibre Channel (FC) are examples of interfaces currently in the standards process. Synchronous Optical Network (SONET) is an example of standardization of higher speeds in the telecom industry. Protocol and software standards have also benefited from committee input.

HIGH-PERFORMANCE PARALLEL INTERFACE (HIPPI)

The HIPPI effort was started by the Los Alamos National Laboratory in early 1987. Our motivation was to have the vendors in the supercomputer community agree on a physical interface standard so that separate interface adapters would not be required to connect to each vendor's proprietary interface. When we took our proposal for an 800 Mbit/s interface to the ANSI Task Group X3T9.3 we were labeled as the "lunatic fringe - who in the world would need anything that fast". Needless to say, we are no longer the "lunatic fringe", in fact some people are saying that we aimed too low.

HIPPI was the first hardware standard in the super-computing arena. You may have heard of HIPPI previously as HSC or HPPI. The name was changed to avoid infringing on existing DEC and Hewlett-Packard trademarks. Some of the initial X3T9.3 goals for HIPPI included:

- a fire hose for moving data at 800 or 1600 Mbit/s,
- get it done quickly since we had immediate needs,
- use current technology - no new silicon required,
- avoid options, and
- keep it simple.

We achieved these goals, and the first HIPPI interfaces were delivered in late 1988. Since then many vendors have implemented HIPPI on their products, or are in the process of implementing HIPPI. Currently HIPPI is the interface of choice in the supercomputing arena.

HIPPI provides a point-to-point simplex data path; that is, it transfers in one direction only. Two back-to-back HIPPIs provide full duplex or dual simplex operation. 800 Mbit/s is supported on one cable, 1600 Mbit/s requires two cables. The cables use twisted-pairs copper wires, are limited to 25 meters in length, and are about 1/2 inch in diameter. Standard ECL drivers and receivers are used.

The hierarchy within HIPPI is:

- Connection - must exist before data can be transferred
- Packet - Groups multiple bursts together into a logical entity
- Burst - Up to 1 or 2 KBytes, basic flow control unit, words within a burst are transferred synchronously with a 25 MHz clock, a checksum follows each burst
- Words - 32 bits on 800 Mbit/s HIPPI, 64 bits on 1600 Mbit/s HIPPI plus an odd parity bit for each byte in each word

HIPPI also provides a flow control mechanism that allows full bandwidth over many kilometers - for use with fiber optic extenders or across other networks such as SONET. Flow control is done on 1 KByte or 2 KByte bursts, decreasing the physical level overhead. Error detection is done in a modular fashion

on individual bytes and bursts; supporting very large (megabyte) packets in a consistent fashion. Error recovery is the responsibility of higher layer protocols.

Networking at the physical layer is supported by HIPPI addressing and "connection" constructs. A common HIPPI network architecture uses a crossbar type circuit switch, for example a Network Systems Corporation PS8 Hub. It works much like your normal telephone connection. That is, the HIPPI source provides a destination address (phone number) and the destination signals whether or not it can accept the connection (answers the phone or hangs up). Once a connection is made, multiple packets of data may be passed without further interaction with the switch, i.e., the only overhead is while the connection is being completed. Either end may hang up, terminating the connection.

The suite of HIPPI documents has expanded beyond the physical layer (HIPPI-PH) described above. HIPPI-SC (Switch Control) defines how physical layer switches operate and are addressed. The HIPPI-FP (Framing Protocol) operates much like a data link layer; breaking large packets up into smaller bursts for transfer across HIPPI-PH, and providing a header describing who the packet belongs to and where the data is located in the packet.

Multiple protocols are supported above HIPPI-FP. HIPPI-LE (802.2 Link Encapsulation) provides a mapping to the IEEE 802.2 data link for support of common network protocols such as TCP/IP. HIPPI-MI (Memory Interface) provides commands for reading and writing memory systems attached via HIPPI A mapping to the Intelligent Peripheral Interface (IPI-3) command sets for disks and tapes is also supported, and is currently being used for stripped disk products.

The status of the HIPPI documents in September of 1991 is:

- HIPPI-PH - an approved ANSI standard
- HIPPI-FP - in public review
- HIPPI-LE - in public review
- HIPPI-MI - just starting the review cycle
- HIPPI-SC - just starting the review cycle

The mapping to IPI-3 will probably be done as revisions to the existing IPI-3 standards rather than a separate HIPPI document. These revisions would also include mappings between IPI-3 and Fibre Channel. The HIPPI-PH document has been submitted to ISO, the International Organization for Standardization, and the other HIPPI documents will also be submitted when they are further along.

FIBRE CHANNEL (FC)

(Yes the name is spelled correctly - the documents will be submitted as international standards, and internationally the spelling is "fibre".)

When the standardization effort for HIPPI started in 1987, ANSI Task Group X3T9.3 wanted to use fiber optics for the increased distance and EMI/RFI benefits. Unfortunately, the fiber optic technology was not mature enough at that time, so HIPPI was based on copper cables to meet the time and simplicity goals. FC is a follow-on to HIPPI, building on many of the concepts introduced with HIPPI. FC is also being developed in ANSI Task Group X3T9.3.

While HIPPI is more of a communications interface, FC was intended to also address the need for a faster I/O channel for supporting peripherals. FC is structured to support the IPI-3 command sets for disk and tape, Small Computer System Interface (SCSI) command sets, IBM S/370 Block Multiplexer commands, and HIPPI-FP packets.

FC, like HIPPI, is also a point-to-point interface, but FC is more general and supports more types of transfers. FC is more of an "all things to all people" type of interface. In the long run, FC will provide more capability than HIPPI, but its generality also produces more complexity, which in turn makes it harder to specify and implement. HIPPI could almost be built with Radio Shack parts, an effective FC implementation will require custom silicon.

Where options were avoided in HIPPI, FC is full of options. For example, FC supports four speeds with data transfer rates of 12.5, 25, 50, and 100 MBytes/s, corresponding to 132, 266, 531, and 1062.5 Mbaud serial signalling rates. The FC media may be single mode fiber or two sizes of multimode fiber, or even inexpensive copper coax cable for short distances. Optical transmitters may be LEDs or lasers. Combinations of the above are specified for different speeds and distances.

HIPPI operates in a datagram mode where higher layer protocols worry about error recovery and retransmission. HIPPI also limits transfers to a single

packet at a time, where the packet may be of any size. In contrast, FC supports three classes of service:

- Class 1 - Dedicated connection, guaranteed delivery, frames received in transmitted order
- Class 2 - Frame switched, buffer-to-buffer flow control, guaranteed delivery, frames may be reordered, virtual connections
- Class 3 - Datagrams, delivery and frame ordering not guaranteed

Class 1 is seen as very useful for visualization, where a dedicated connection may exist for long periods of time, and interference from other data streams is undesirable. Class 2 will probably be used heavily for traditional I/O transfers, where multiple transfers are open at one time with frames from the different transfers multiplexed on a single fiber. Class 3 can be used with traditional communications protocols where recovery and re-ordering are already handled in the upper layer protocols, and where connection set-up times must be avoided.

FC is structured into four layers for ease of understanding and documentation. FC-0 specifies the physical layer with the serial drivers, receivers, media, etc. FC-1 specifies the 8B/10B encoding/decoding scheme used to encode the data into a DC balanced serial bit stream. FC-1 also defines special symbols for such things as Idle, SOF, EOF, etc. FC-2 defines the framing, e.g., where the address, control, data, and check fields are located and what they mean. FC-3 defines common services such as striping a single packet across multiple FC-0's for higher bandwidth, hunt groups, and multicasting. FC-4s are the mappings to higher layer protocols, e.g., to the IPI-3 command sets for disk and tape.

The logical hierarchy within FC is:

- Operation - Logical construct to identify and group things for an upper layer protocol
- Exchange - Group of sequences, normally related to I/O control blocks
- Sequence - Unidirectional group of frames
- Frame - Basic transfer unit, contains header with addresses, control, offsets, etc., contains up to 2 KBytes of data, basic flow control unit, contains checksum, words within a frame are synchronous

Identifier and offset fields are contained within each frame's header, allowing the receiving port to place the data in the proper place in memory, hopefully eliminating the need for data copies in the receiving computer. Considerable work has gone into providing multiple levels of indirection so that the individual frames can be disposed of by state machines implemented in silicon rather than having to be handled by a general purpose processor. The feeling is that this is mandatory if we are to keep up with the data transfer rate, multiplexed frames, and the variety of applications.

NETWORK ARCHITECTURES

HIPPI and FC provide point-to-point connections which can be used as the basic building blocks for computer networks. Different types of network architectures are appropriate for different applications. HIPPI and FC lend themselves to ring and circuit switch architectures.

Circuit switch architectures

For comparison, circuit switching is what is used in the telephone system today. That is, your call is separate and independent from someone else's call, even though you are both using the same circuit switch hardware. The separate but independent nature of circuit switching is one of the requirements for visualization. The Los Alamos National Laboratory is prototyping a circuit switching architecture called the Multiple Crossbar Network.

Figure 1 shows a 4 x 4 crossbar switch interconnecting four hosts. Note that connections exist for simultaneous transfers from Host 2 to Host 4, and from Host 3 to Host 1. The "CBI" nodes are "CrossBar Interfaces", in the Los Alamos nomenclature. They would perform such functions as data buffering, switch access, address resolution, security checking, and low level protocols. The CBIs are very similar to the CABs for the Carnegie Mellon NECTAR project being developed by Network Systems.

The circuit switch components run at the basic channel rate, and obtain a high total bandwidth by allowing multiple channels to be active simultaneously. For example, an 8 x 8 circuit switch for HIPPI would have each channel running a 800 Mbit/s, the circuits within the switch running at 800 Mbit/s, and a total bandwidth of 6400 Mbit/s. In use, one mainframe may be sending data to a visualization station, while another mainframe is reading data from a disk system, with

both simultaneously transferring data at 800 Mbit/s rates.

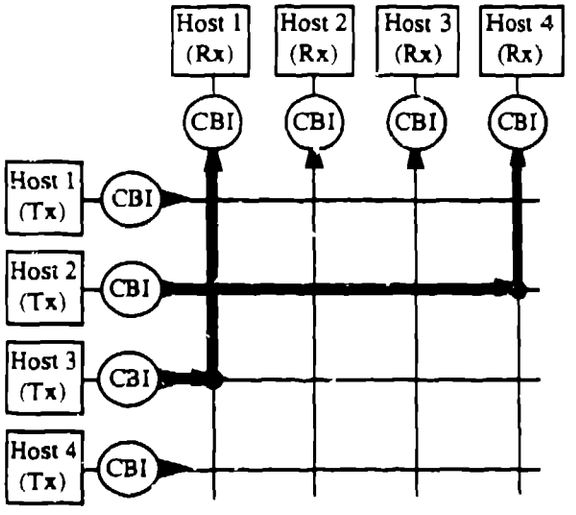

Figure 1. Circuit switch architecture

Normally, once a connection is completed, the channel operates as if there were no switch involved. That is, delays may occur on circuit setup, but no delays, other than circuit delays, are encountered once the connection is completed.

Circuit switches utilize different access control mechanisms from traditional bus or ring architectures. Namely, if a source on a switch finds that its requested destination is busy, and if the source has data for a different destination, then the source can try sending to the second destination. With a bus or ring, if the media was busy, you could not send even if you had data for another destination.

Camp-on features may also be used to hang a source waiting for a specific destination to complete. Call queueing schemes have also been proposed for connection setups. Switch systems need to watch out for hung channels and channel hogs.

In the absence of a busy destination, setting up a circuit may take from a microsecond to a millisecond, depending on the switch size and connection control circuitry. Once completed, delays through the switch from a few nanoseconds to a microsecond may be encountered.

While a ring or bus system may grow indefinitely one attachment at a time, circuit switches grow in major increments. For example, if you are using an 8 x 8 switch and want to add a ninth element, then you have to buy another whole 8 x 8 switch and interconnect the switches. Switch architectures are often square, e.g., crossbars, but may be tailored to a variety of applications. For example, a local switch may interconnect several workstations but have only one connection to the main switch; supporting only one mainframe to workstation transfer at a time.

There are advantages to large switches, e.g., up to 4096 connections, and to small modular switches, e.g., 8 x 8 or 32 x 32, and vendors are building both. Some of the early uses may give us some guidelines on the best way to apply switches.

Ring architectures

Ring networks provide a single data path that is shared by all of the attachments. This single data path limits the total bandwidth, but does give a natural broadcast capability. Bus access is usually determined by token passing or time slots. An advantage of rings is that it is usually fairly easy to add one more station. FDDI is an example of a ring network running at 100 Mbit/s.

Figure 2 shows a ring network interconnecting four hosts. The "RI" elements are "ring interfaces" for performing such functions as data buffering, ring access, data buffering, security checking, and low level protocols.

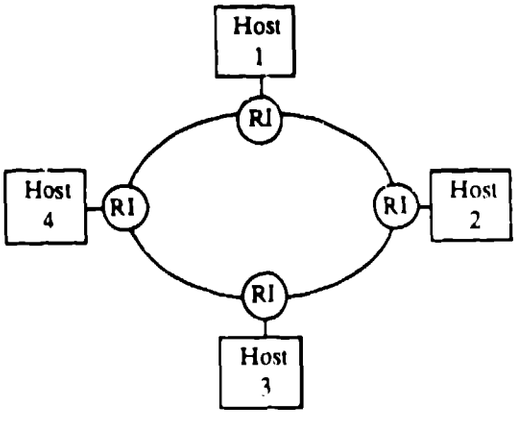

Figure 2. Ring architecture

FC based rings are being considered for connecting peripherals, e.g., disks, to mainframes. In this environment, the limitation of a single data path is not critical since the mainframe is normally the single generator and user of the data. It is envisioned that these rings would be cheaper than a circuit switch architecture.

Wavelength Division Multiplexing (WDM)

Wavelength division multiplexing (WDM) operates by sending multiple data streams, each at a separate wavelength (i.e., frequency), on a single fiber. For comparison, FC uses baseband signalling, sending only a single stream down a fiber.

WDM can be compared to the lead-in cable for your TV set; there is only one cable, but there are multiple station's signals on that cable. Figure 3 shows one version of a WDM network interconnecting four hosts. In figure 3, each host transmits on a fixed wavelength, λ1 through λ4. At each receiver, the tunable filter selects the appropriate wavelength to listen to a speci transmitter. Another version of a WDM netw would have each receiver set to a unique single wavelength, and the transmitters tune to the different wavelengths. Still another version would have both the transmitters and receivers tunable. The network can theoretically have a very large number of channels, e.g., 2500 channels, each 1 GHz with 9 GHz guard bands. This is based on a center wavelength of 1.55 nm and tuning from 1.45 nm to 1.65 nm.

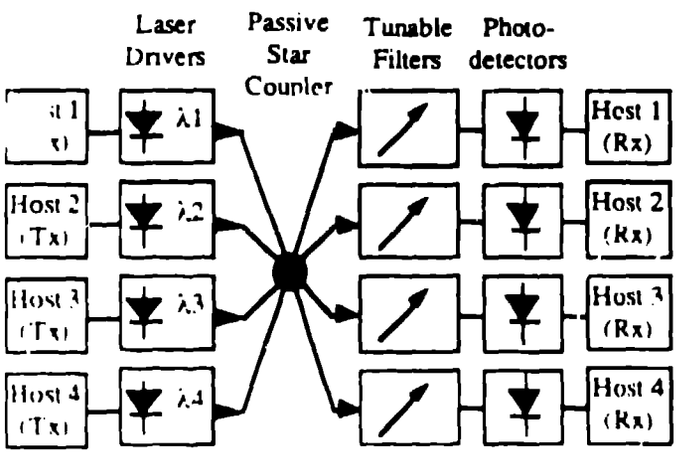

Figure 3. WDM network example

There are some basic problems that need to be solved before WDM becomes practical for computer networks. The tuning needs to be fast (less than 1 microsecond) and accurate (to get the maximum number of channels). There also needs to be minimum crosstalk (for the maximum number of channels and adequate bit error rate). There also need to be in-line broadband amplifiers to overcome the losses of the star couplers. Ways to distribute the star coupler to the end points would also help. And of course, the parts need to be inexpensive and mass producible (hand selection of laser wavelengths is not acceptable).

Todays computer networks use what is called "in-band addressing" i.e., the destination address is carried along with e message, not routed on a separate control path. Also, most of todays computer networks use packet switching with datagrams as the underlying transfer mechanism. Here each message is a separate entity with addressing and error control portions. Rather than using packet switching, WDM networks seem to lend themselves more towards circuit switching. With circuit switching a pilot message is sent out to establish a path (circuit) between the source and destination. Once the path is established, the data message can then be transmitted. This circuit set-up adds to the message latency.

With WDM the path must be established, i.e., both the transmitter and receiver must be using the same wavelength, before the message packet can be sent. If you are using tunable receivers and fixed transmitters, then how does the receiver know when a transmitter wants to send something to it so that the receiver can tune to the transmitter's wavelength? Likewise, if the receiver is fixed and the transmitter tunable, then how does the transmitter know that someone else isn't already transmitting on that wavelength? If someone else is transmitting on this wavelength, then the messages will collide resulting in neither message getting through correctly. There are ways to solve this "media access" problem, but most of them require some sort of "out-of-band addressing" at different wavelengths. The problem is not insurmountable, it is just a problem. This media access problem will affect the latency from source to destination. With changes to accommodate the access differences, FC should work well with WDM.

Interfaces to the telecommunications world

The telcom networks and computer networks have traditionally used different techniques. The telcom networks have effectively used circuit switching and

time division multiplexing of many slow channels to a single fast channel. The computer networks have used packet switching with datagrams, where each packet takes the total bandwidth of the media. The telcom networks have been very concerned with guaranteed bandwidth so that the data is not delayed, for example causing uneven time delays in speech traffic. The computer networks were less worried about incremental delay, and were more concerned with making use of all of the available bandwidth.

Now we are seeing the two "cultures" starting to merge. The computer networks need some of the guaranteed bandwidth circuit switching techniques to transmit video and voice among the end nodes. Likewise, the telcom networks are becoming digital and using small packets, e.g., 53-byte cells in Asynchronous Transfer Mode (ATM) of the Synchronous Optical Network (SONET), for carrying multiple traffic streams. The telcom networks still need a call set-up to load the address translation look-up tables in the route.

Conventional wisdom says that the less you 'touch' a packet, the lower the overhead. That is, an interface or bridge that can 'touch', or operate on, 20,000 packets per second is a real screamer, and effectively takes 50 microseconds for each packet. At 800 Mbit/s, 50 microseconds translates into a 5 KByte packet. At the 2.4 Gbit/s speed of SONET, a 53-byte packet takes less than 200 nanoseconds, hence assembling and working with 53-byte cells is going to be a challenge at the higher SONET rates, e.g., approximately 5,600,000 cells per second.

Other potential problems associated with ATM include the fact that the cells do not include any error detection, e.g., parity, on the data portion of the cell. Cells may also be discarded by intermediate switches during overload conditions.

OPEN PROBLEMS REQUIRING FUTURE WORK

HIPPI and FC may be the lower layers of future network architectures. With these higher speed physical connections, there is incentive to work on the next bottleneck, which may well be the Transport Layer. TCP/IP and TP4 are the most widely used transport layers, but they may not perform well in the gigabit environment.

Existing upper layer protocols were designed to operate with yesterdays physical layers. Now, rather than error rates of 10^{-4}, error rates of 10^{-9} are expected, largely due to the improvements from using fiber optic components. The distances and transfer rates also affect the protocol. The delay between California and New York is 30 milliseconds, allowing 3000 packets of 1 KBytes each to be in transit. Window sizes, flow control, and error recovery at the higher speeds need to be addressed.

Supercomputers have proven to be very effective for simulating physical phenomenon. Congress, in an attempt to increase the effectiveness of the United States, is pushing a National Research and Education Network (NREN), with a goal of a coast-to-coast 3000 Mbit/s computer network backbone. If you cannot move the users to the computers, then make the computers available to the users as if they were adjacent. There is a lot of research and testing going on to make the NREN a reality within the time frame goal. Los Alamos is participating in the Casa testbed. HIPPI is also being used heavily in the testbeds.

Interoperability with the telephone switching systems is required to realize the NREN. The telecom industry has been promoting Asynchronous Transfer Mode (ATM) for switching and routing. ATM uses a basic cell size of 48 bytes plus a 5-byte header. ATM makes good sense when supporting many voice circuits, how well it works with gigabit/s data transfers remains to be seen.

SUMMARY

Computer networks operating at gigabit per second transfer rates are seen as necessary for many applications, and gigabit networks are becoming available. HIPPI and FC will provide some of the basic building blocks for these networks. Further work needs to be done in higher layer protocols, and long distance networks, to achieve our national goals.

ACKNOWLEDGEMENTS

The Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405 ENG-36. This work was performed under auspices of the U.S. Department of Energy.